

A complex network graph with numerous nodes of varying sizes and colors (shades of blue and grey) connected by thin lines, set against a dark grey background. The nodes are distributed across the frame, with a higher density in the center.

Quantitative Approaches to Political Science Research

Quantitative Approaches

- We've covered:
 - Mean, median, mode
 - Variance, standard deviation, standard error
 - χ^2 tests, t-tests, and Pearson's correlation coefficients
- Today we'll cover:
 - Bivariate and multivariate linear regression
 - Time permitting, maximum likelihood approaches (logit/probit)

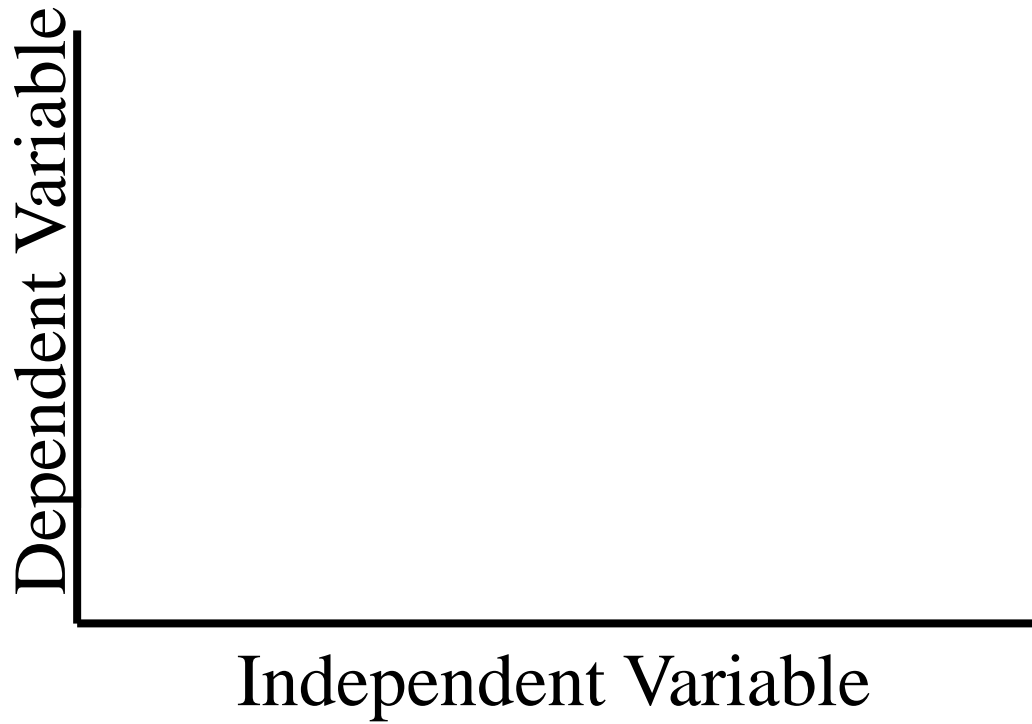
Quantitative Approaches

Table 8.1 Variable types and appropriate bivariate hypothesis tests

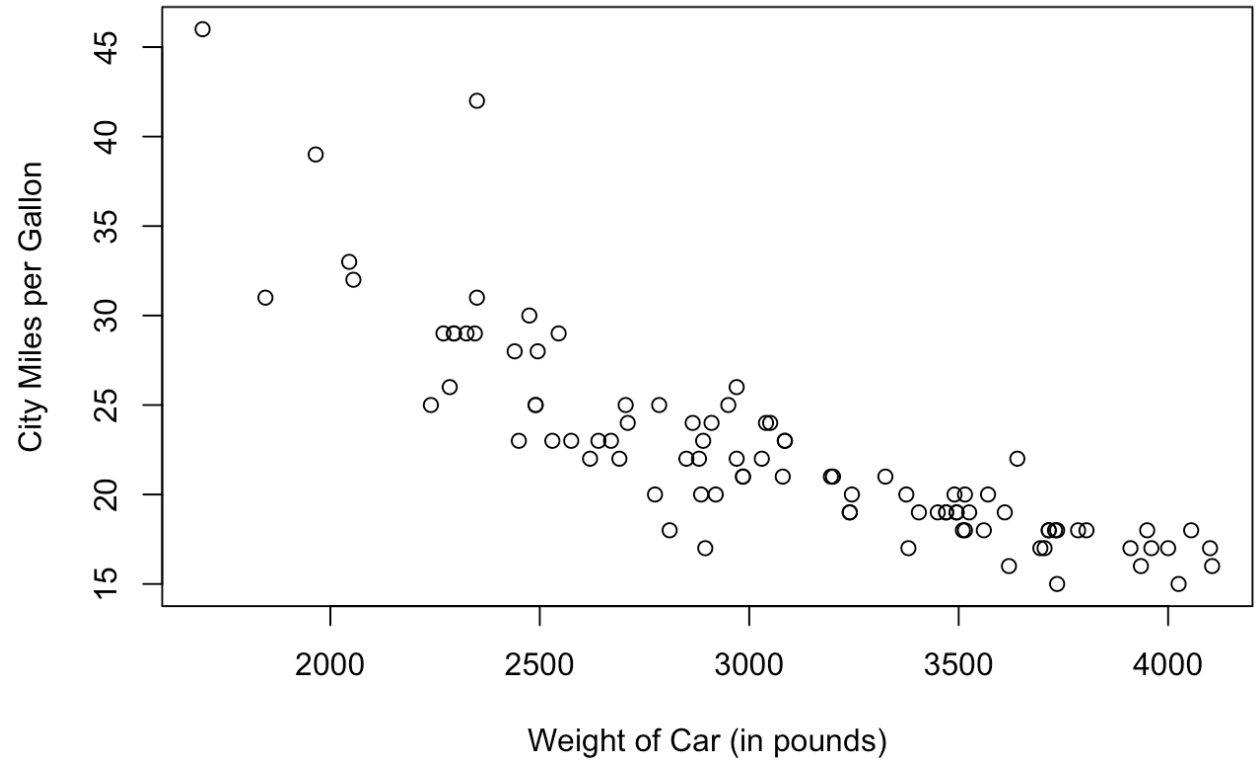
		Independent variable type	
		Categorical	Continuous
Dependent variable type	Categorical	<i>tabular analysis</i>	probit/logit (Ch. 12)
	Continuous	<i>difference of means;</i> regression extensions (Ch. 11)	<i>correlation coefficient;</i> two-variable regression model (Ch. 9)

Note: Tests in italics are discussed in this chapter.

Linear Regression

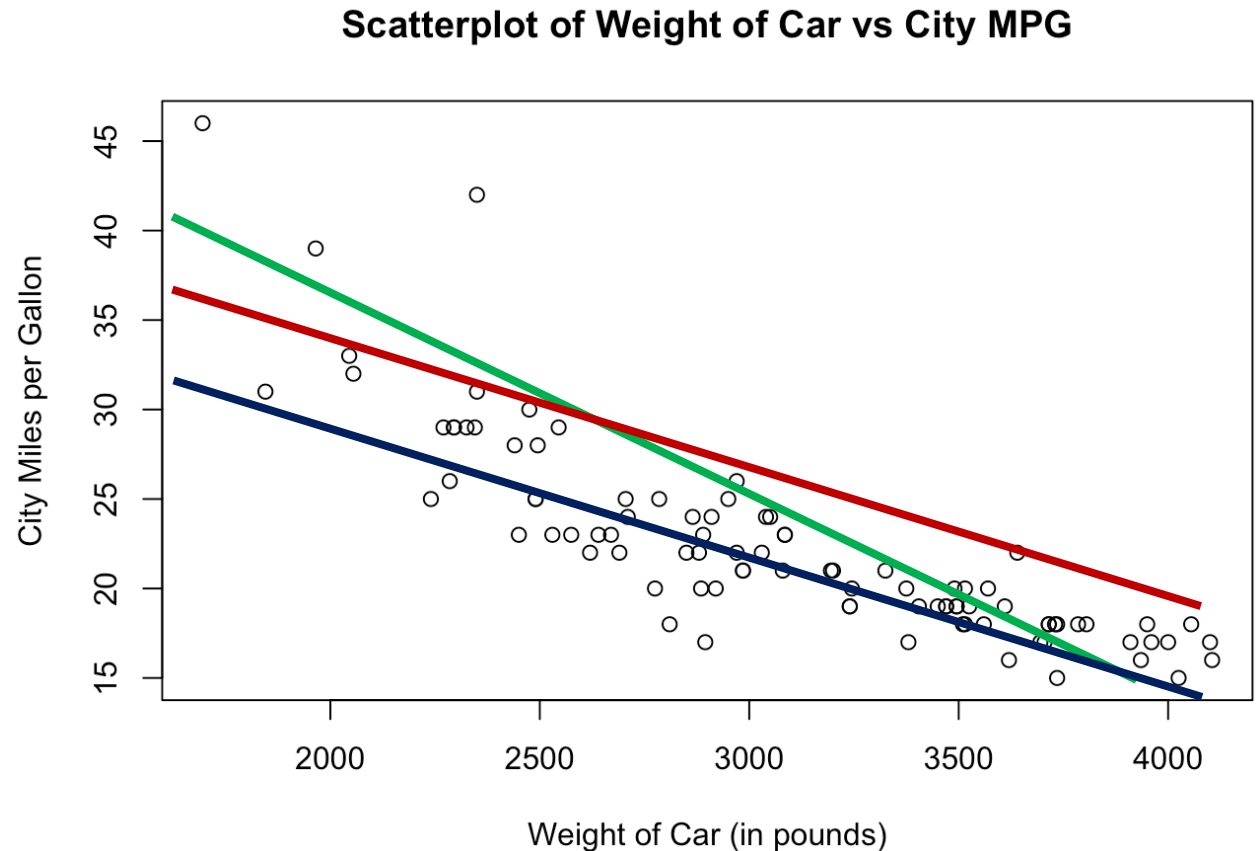


Scatterplot of Weight of Car vs City MPG



Linear Regression

- For linear relationships...
- $y=mx+b$
- Where to fit this line?
- We require a systematic means to find the line of best fit
 - Least squares approach



Linear Regression

- We need to move past the $y=mx+b$ formulation

$$y_i = \alpha + \beta x_i + u_i$$

- y_i : dependent variable
- α : Y intercept
- β : slope coefficient
- x_i : independent variable
- u_i : residual

Linear Regression

- We need to move past the $y=mx+b$ formulation

$$y_i = \alpha + \beta x_i + u_i$$

- y_i : dependent variable
- x_i : independent variable
- α : Y intercept

Linear Regression

- We need to move past the $y=mx+b$ formulation

$$y_i = \alpha + \beta x_i + u_i$$

- β : slope coefficient
 - Consider this as the effect of your independent variable on the dependent variable ($Y=\beta * X$)
- u_i : residual
 - Residual here is synonymous with error; this is the degree of deviation from the line of best fit and the observed value

Linear Regression

- We need to move past the $y=mx+b$ formulation

$$y_i = \alpha + \beta x_i + u_i$$

- We also need to be conscious that this is the population equation
- This is the data generating process (DGP) in the world, not necessarily in our sample

Linear Regression

- We need to move past the $y=mx+b$ formulation

$$y_i = \hat{\alpha} + \hat{\beta}x_i + \hat{u}_i$$

- We also need to be conscious that this is the population equation
- This is the data generating process (DGP) in the world, not necessarily in our sample
- When are discussing estimated values, we put a little hat on the variables – alpha hat, beta hat, etc.
- This is the sample regression equation

Linear Regression

- We need to move past the $y=mx+b$ formulation

$$y_i = \hat{\alpha} + \hat{\beta}x_i + \hat{u}_i$$

- Because the α , β , and \hat{u} values are not *known*, but rather *estimated*, they get hats
- This signals that we don't know these values, and we likely cannot know these values, but we estimate a range of likely values within which the population ('true') value lies – recall confidence intervals

Linear Regression

$$y_i = \hat{\alpha} + \hat{\beta}x_i + \hat{u}_i$$

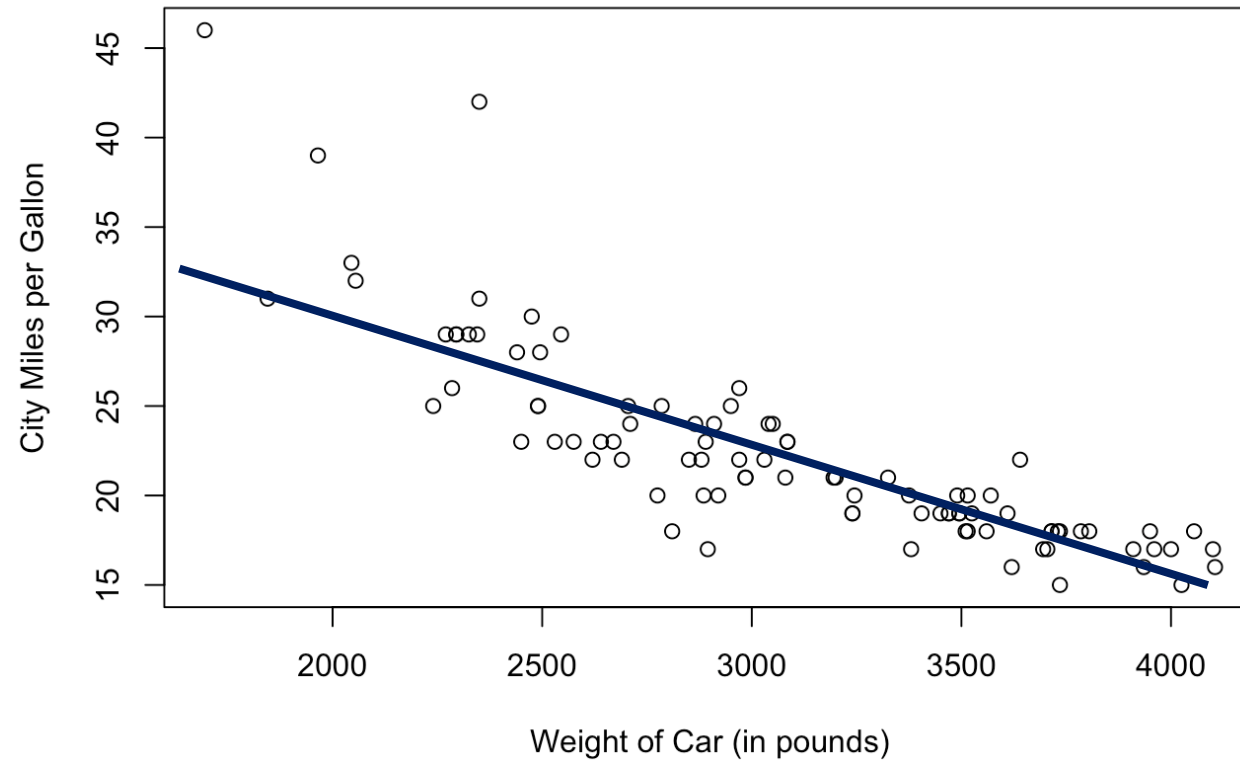
- Recall the difference between stochastic and systematic variation
- Both operate in this equation
- We can dichotomize \hat{u}_i into two components

$$\hat{u}_i = \varepsilon_i + u_i$$

- Where \hat{u}_i are our observed residuals, ε_i is the unmodelled systematic variation, and u_i is the remaining random (stochastic) variation

Linear Regression

Scatterplot of Weight of Car vs City MPG



Linear Regression

$$y_i = \hat{\alpha} + \hat{\beta}x_i + \hat{u}_i$$

- Recall the difference between stochastic and systematic variation
- Both operate in this equation
- We can dichotomize \hat{u}_i into two components

$$\hat{u}_i = \varepsilon_i + u_i$$

- We will always have some level of \hat{u}_i due to both stochastic and systematic variation, no matter how many IVs we add

Linear Regression

$$y_i = \hat{\alpha} + \hat{\beta}x_i + \hat{u}_i$$

- Recall the difference between stochastic and systematic variation
- Both operate in this equation
- We can dichotomize \hat{u}_i into two components

$$\hat{u}_i = \varepsilon_i + u_i$$

- As we'll see later, it is not always beneficial to add more IVs to decrease \hat{u}_i

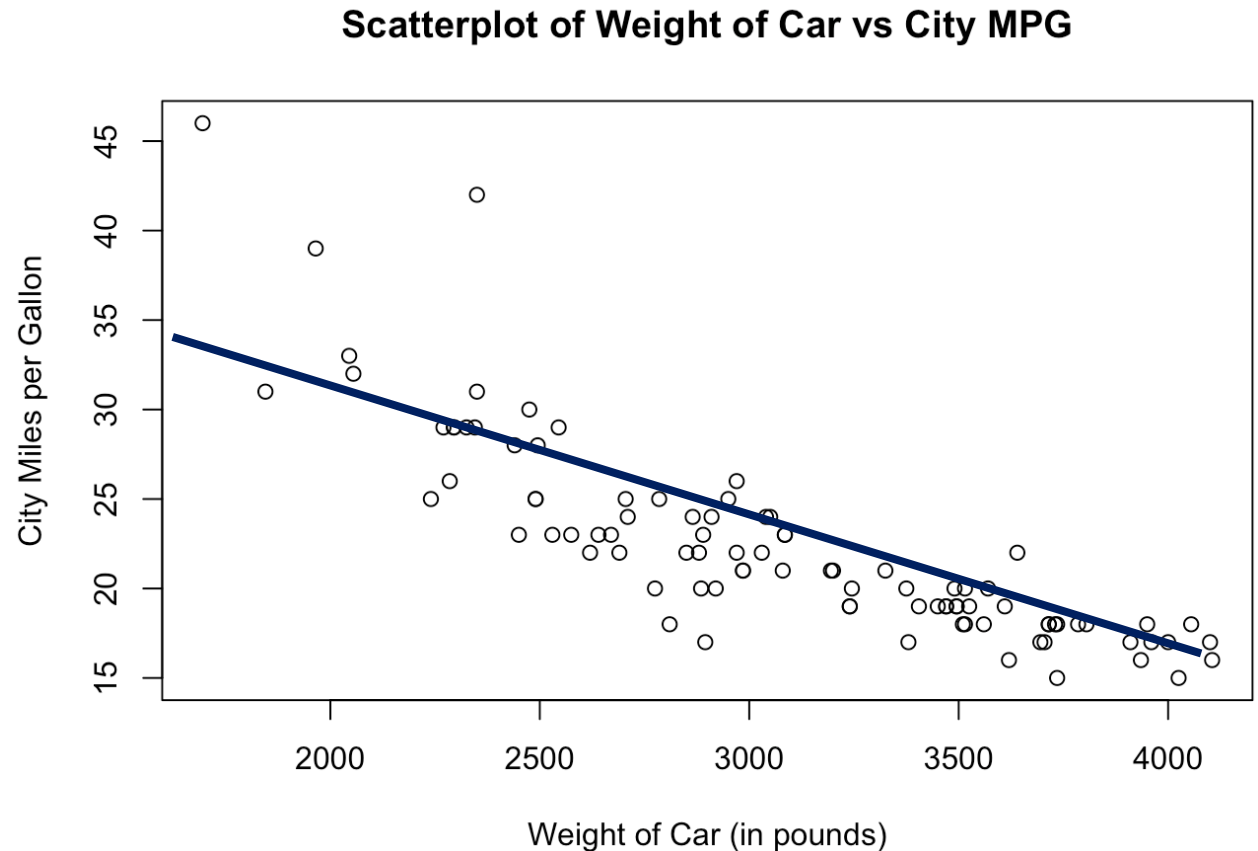
Linear Regression

$$y_i = \hat{\alpha} + \hat{\beta}x_i + \hat{u}_i$$

- What do we really care about in this equation?
- We want to find $\hat{\beta}$ as this is the estimated effect of our independent variable(s) on the observed outcome (DV)
- But we also need to know where this effect begins ($\hat{\alpha}$) as well as our degree of confidence about our estimates (\hat{u}_i)

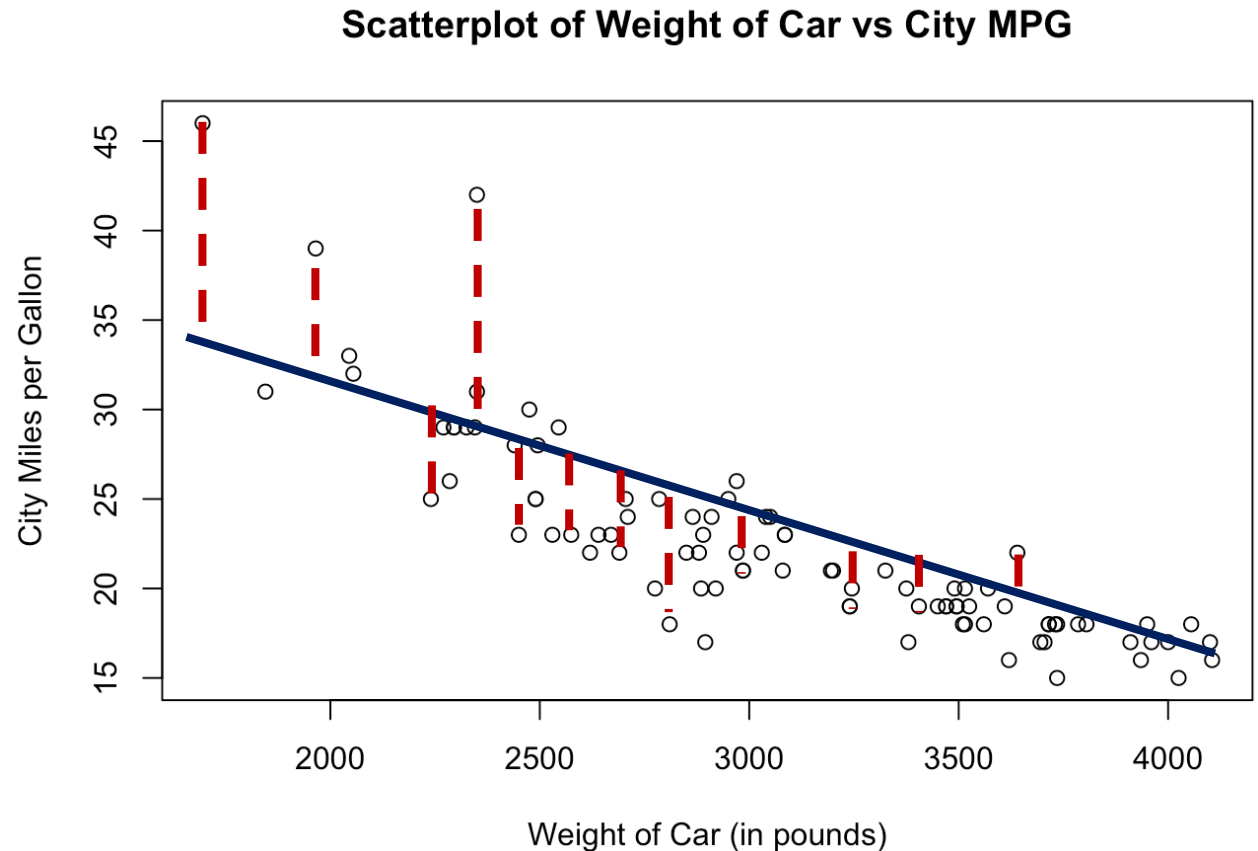
Linear Regression

- Returning to the scatterplot
- We want to find the line of best fit
- This is determined by the line that minimizes distance between our observations and the linear line



Linear Regression

- Returning to the scatterplot
- We want to find the line of best fit
- This is determined by the line that minimizes distance between our observations and the linear line



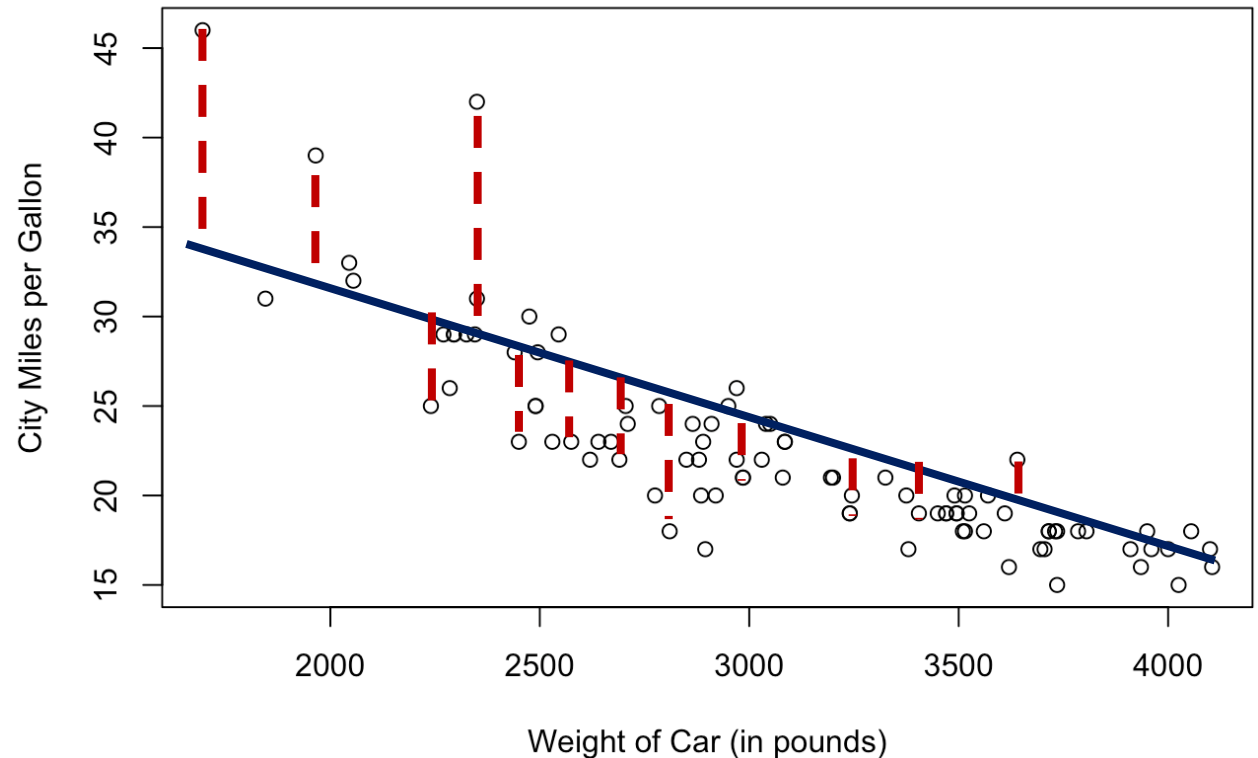
Linear Regression

- This is performed via the two equations below

- $$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

- $$\alpha = \bar{y} - \hat{\beta}\bar{x}$$

Scatterplot of Weight of Car vs City MPG



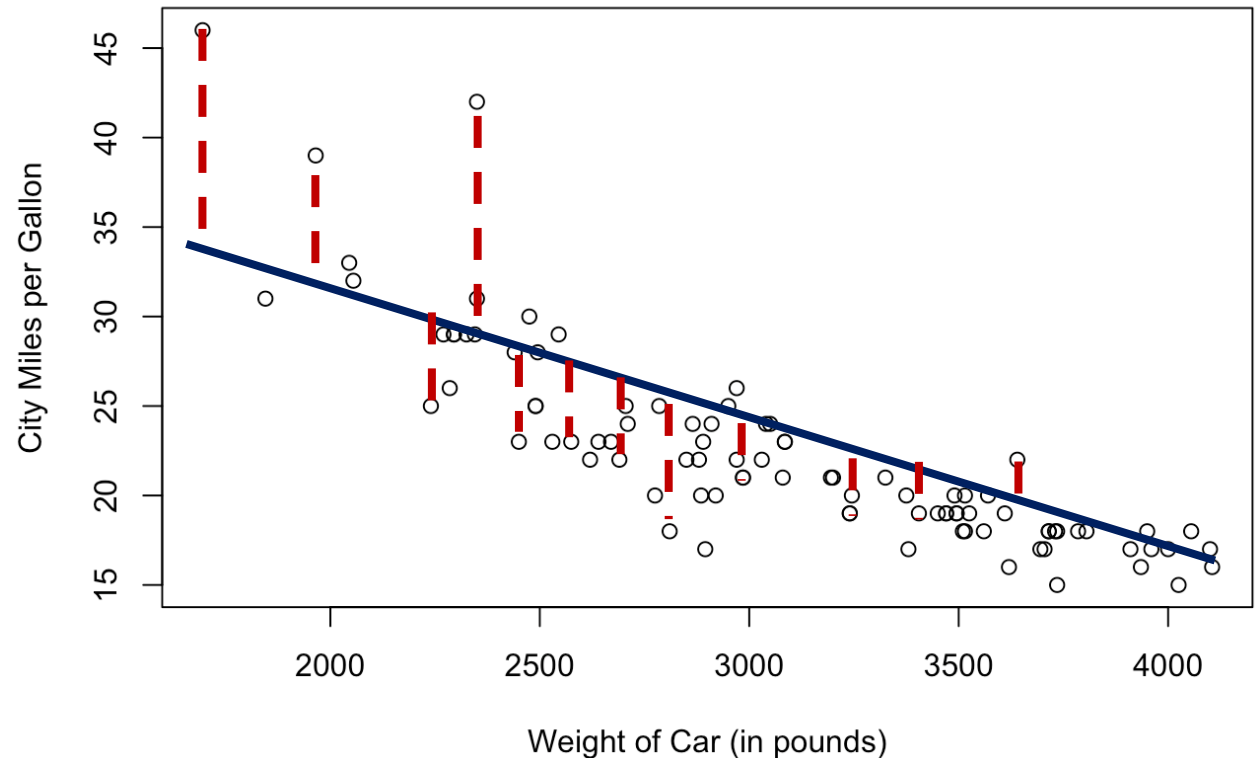
Linear Regression

- This is performed via the two equations below

- $$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})}$$

- $$\alpha = \bar{y} - \hat{\beta}\bar{x}$$

Scatterplot of Weight of Car vs City MPG



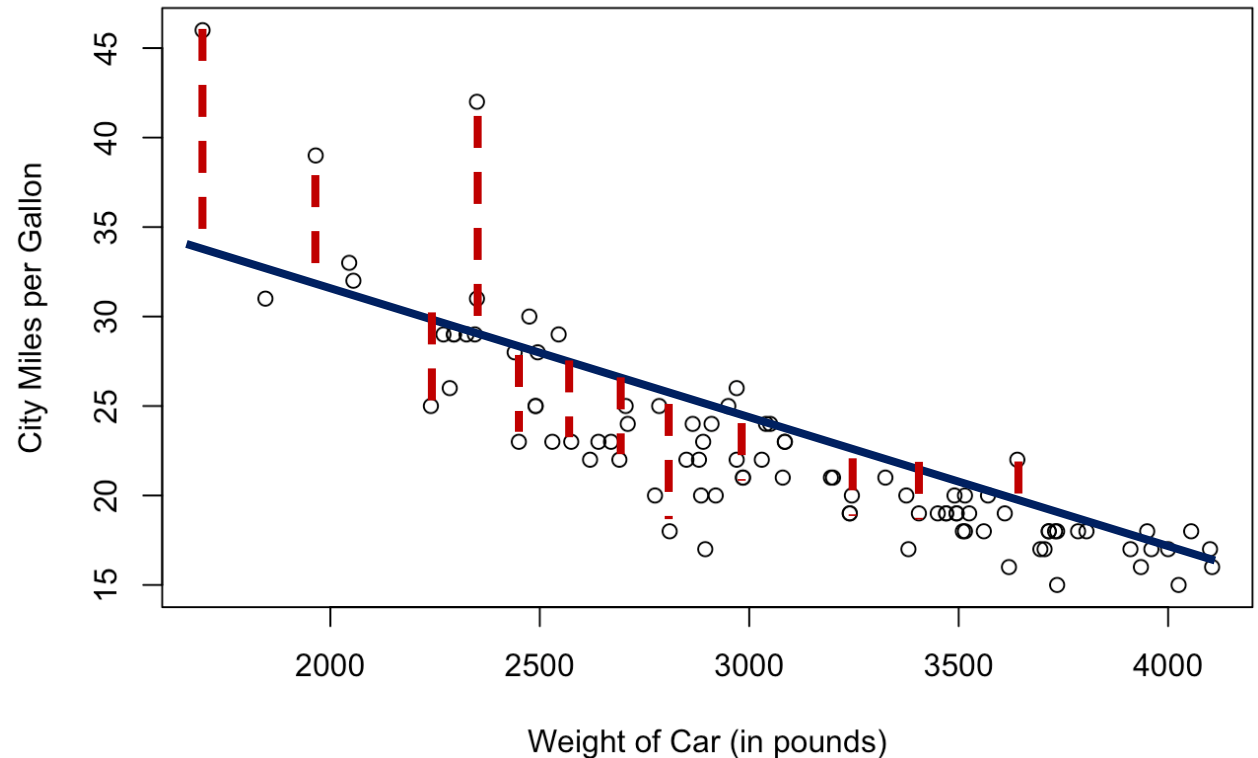
Linear Regression

- This is performed via the two equations below

- $$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})}$$

- $$\alpha = \bar{y} - \hat{\beta}\bar{x}$$

Scatterplot of Weight of Car vs City MPG



Linear Regression

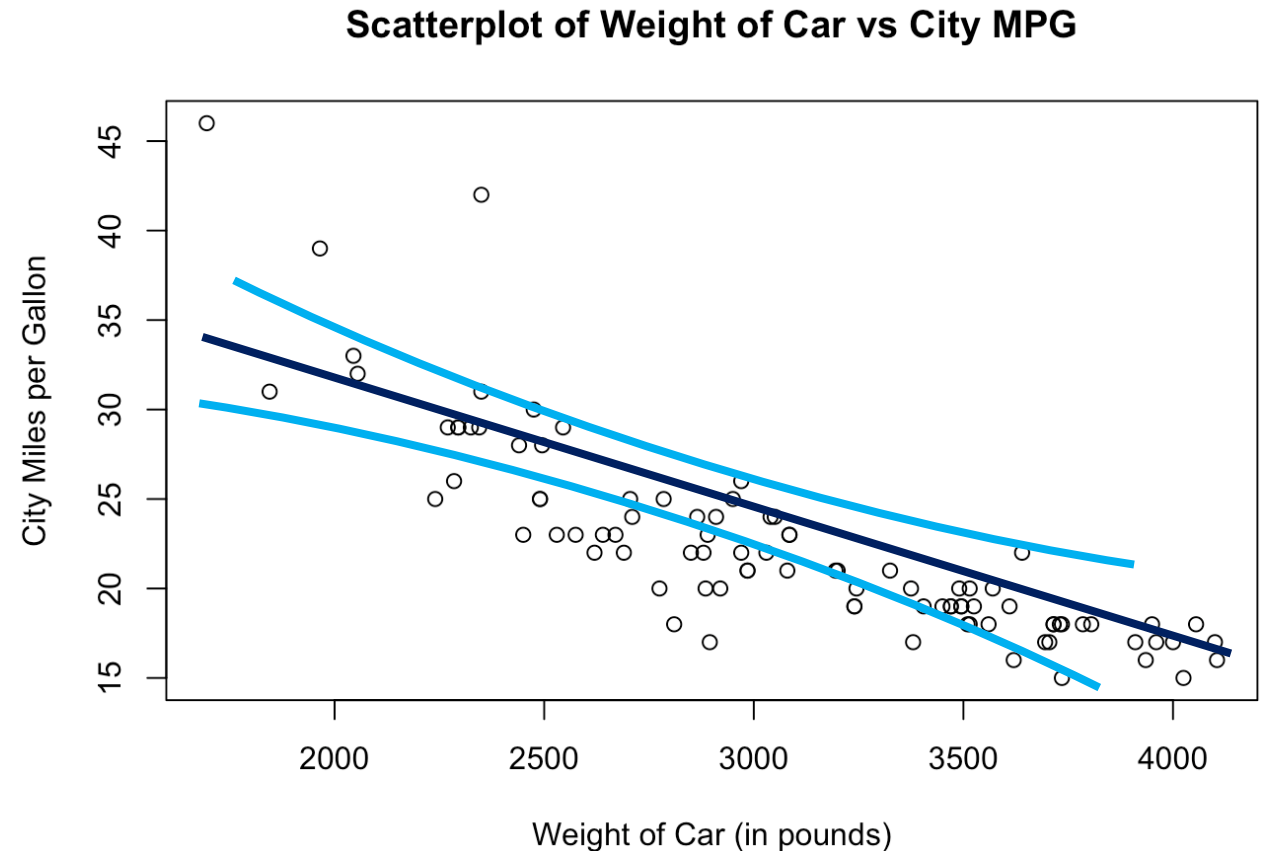
- Once we have our coefficient estimates, we calculate the standard error of the β and α coefficients through the following equations

$$\widehat{SE}(\hat{\beta}) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

$$\hat{\sigma}^2 = \frac{1}{n - k - 1} \cdot \sum_{i=1}^n \hat{u}_i^2$$

Linear Regression

- In doing, we can construct a confidence interval about our regression estimate
- Generally, 95%
- We are 95% confident that the ‘true’ effect of X on Y falls within this range – both slope and intercept



Linear Regression

- This is called *Ordinary Least Squares (OLS)*
- The name derives from the process of finding the line that minimizes the square distances between the regression line and the observations
- This is also termed linear regression or least squares regression
- This process has several benefits as well as a number of restrictions

Linear Regression: BLUE

- We use OLS because it is BLUE
 - The best, linear, unbiased estimator
- Best: Minimum variance between β and $\hat{\beta}$, and α and $\hat{\alpha}$, as the sample size approaches ∞
- Linear: Where the relationship under study is linear, we use a linear estimator
- Unbiased: Accurately estimates the regression coefficients $(\hat{\alpha}, \hat{\beta})$

Linear Regression: Gauss Markov Assumptions

- OLS has a number of assumptions/requirements
- These are known as the Gauss-Markov Assumptions
- The relationship under study must be linear in the population

Linear Regression: Gauss Markov Assumptions

- OLS has a number of assumptions/requirements
- These are known as the Gauss-Markov Assumptions
- Our data is randomly drawn from the population

Linear Regression: Gauss Markov Assumptions

- OLS has a number of assumptions/requirements
- These are known as the Gauss-Markov Assumptions
- The IVs are not perfectly correlated with one another
 - Non-collinearity

Linear Regression: Gauss Markov Assumptions

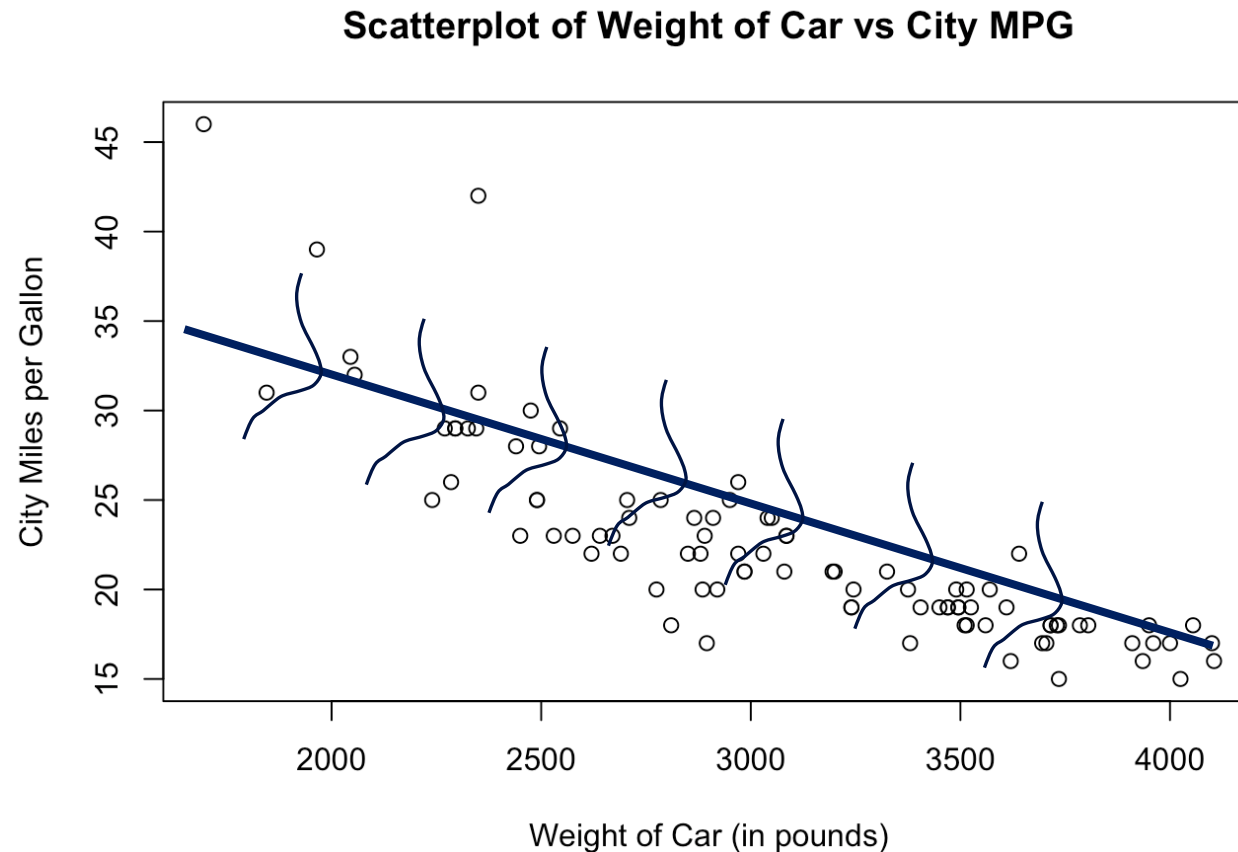
- OLS has a number of assumptions/requirements
- These are known as the Gauss-Markov Assumptions
- The IVs are not correlated with the error term/residuals

Linear Regression: Gauss Markov Assumptions

- OLS has a number of assumptions/requirements
- These are known as the Gauss-Markov Assumptions
- The errors (residuals) are uncorrelated with each other, and the IVs, and with an expected value of 0
- $\text{cov}(u_i, u_j) = 0$
- $E(u_i) = 0$

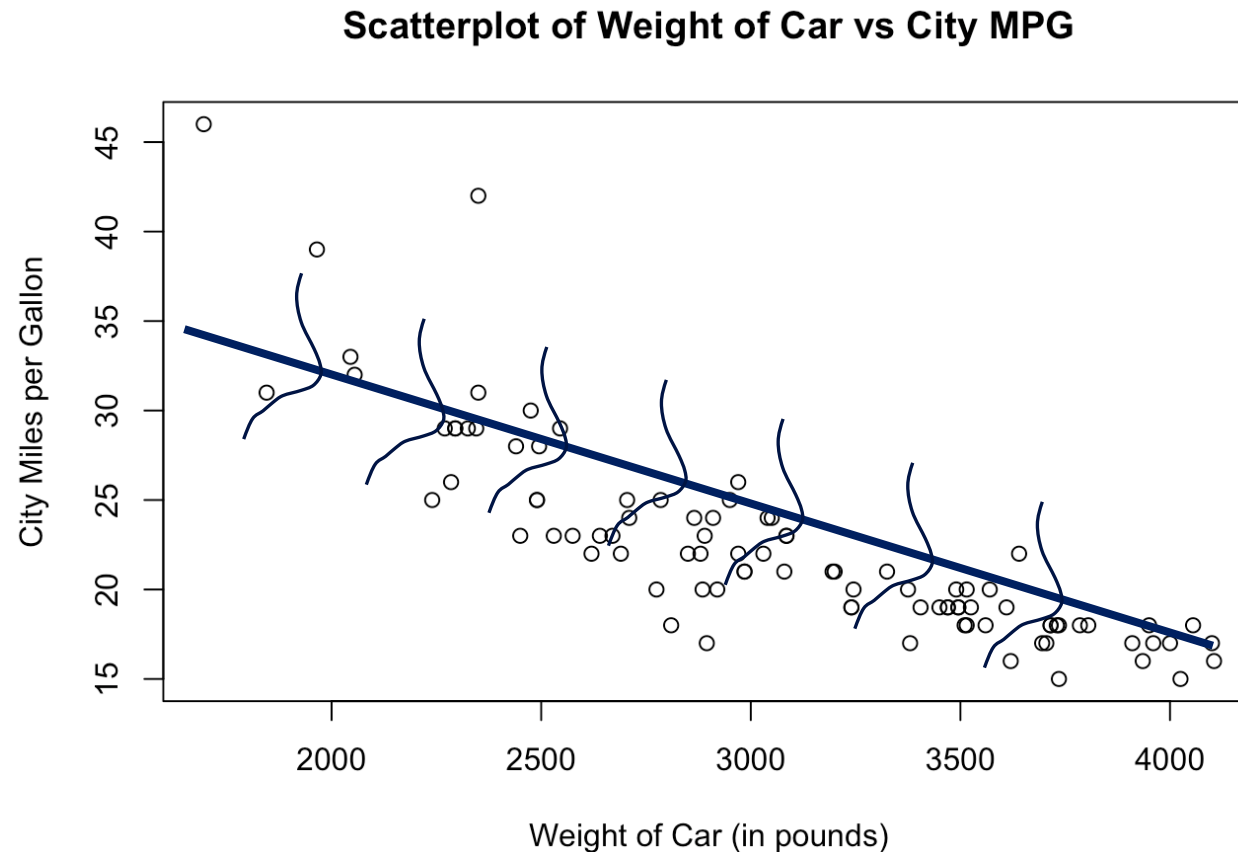
Linear Regression: Gauss Markov Assumptions

- ‘Spherical’ errors
- The errors should be normally distributed about the regression line
- If skewed, this means that the regression line is not ‘splitting’ the data, and is thus biased



Linear Regression: Gauss Markov Assumptions

- ‘Spherical’ errors
- With more than a single IV, we need to conceptualize this in three dimensions
- A circle in three dimensions is a sphere
 - 68/95/99 of the error distribution

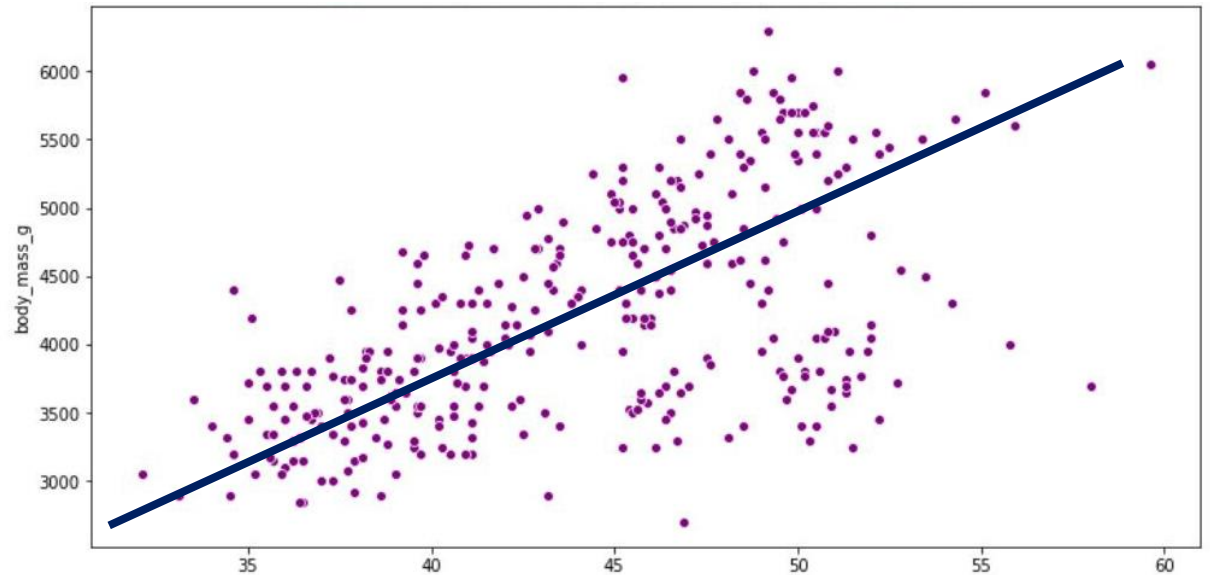
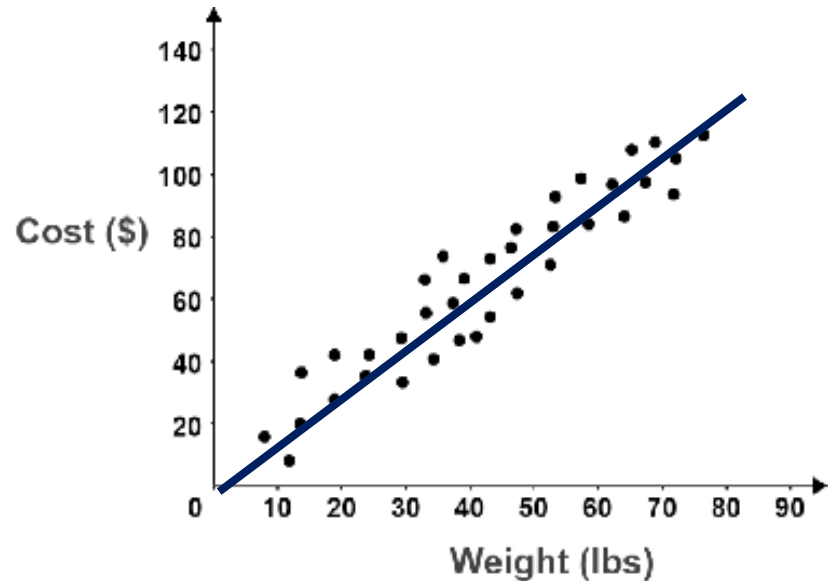


Linear Regression: R^2

- How do we know if the linear regression line is doing a ‘good’ job in predicting the observed data
- Such a measure is inherently contingent on the dispersion of the observed data

Linear Regression: R^2

- How do we know if the linear regression line is doing a ‘good’ job in predicting the observed data
- Such a measure is inherently contingent on the dispersion of the observed data – the variance of the data



Linear Regression: R^2

- How do we know if the linear regression line is doing a ‘good’ job in predicting the observed data
- Such a measure is inherently contingent on the dispersion of the observed data
- Where there is a greater degree of stochastic and systematic variation at work in the observed data, the linear regression estimator will do the best it can
- We can only reduce such variation to a limited degree

Linear Regression: R^2

- How do we know if the linear regression line is doing a ‘good’ job in predicting the observed data
- We quantify the degree of variation explained by the linear regression process by the metric of R^2 and adjusted R^2
- This is also termed the coefficient of determination or “goodness of fit” measure

Linear Regression: R^2

- Total Sum of Squares: the total variation in Y_i
- Residual Sum of Squares: the variation in Y_i not explained by X_i
- $R^2 = 1 - \frac{RSS}{TSS}$

Linear Regression: R^2

- Total Sum of Squares: $TSS = \sum_{i=1}^n (y_i - \bar{y})^2$
- Residual Sum of Squares: $RSS = \sum_{i=1}^n (\hat{u}_i^2)$
- $R^2 = 1 - \frac{RSS}{TSS}$

Linear Regression: R^2

- As we can see from these equations, there is no way to quantify how many independent variables are being used
 - If you add more independent variables, you will explain more of the observed variation
- $TSS = \sum_{i=1}^n (y_i - \bar{y})^2$
- $RSS = \sum_{i=1}^n (\hat{u}_i^2)$
- $R^2 = 1 - \frac{RSS}{TSS}$

Linear Regression: R^2

- As we can see from these equations, there is no way to quantify how many independent variables are being used
 - If you add more independent variables, you will explain more of the observed variation
- Thus, we prefer to use adjusted R^2
- This weights our measure to account for the number of IVs we're using

$$R_{adj}^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1}$$

where :

R^2 = *R - squared*

n = *number of samples/rows in the data set*

p = *number of predictors/features*

Linear Regression: Significance

- We've covered how to calculate the linear regression estimates, how uncertainty is modeled, and how well the model explains observed variation in the outcome variable
- What about statistical significance?
- We calculate a t-statistic by comparing observed values to the value posited by the null hypothesis, over the standard error of the model

$$t_{n-k} = \frac{\hat{\beta} - \beta^*}{se(\hat{\beta})}$$

Linear Regression: Significance

- We've covered how to calculate the linear regression estimates, how uncertainty is modeled, and how well the model explains observed variation in the outcome variable
- What about statistical significance?
- We calculate a t-statistic by comparing observed values to the value posited by the null hypothesis, over the standard error of the model

$$t_{n-k} = \frac{\hat{\beta} - 0}{se(\hat{\beta})}$$

Linear Regression: Tables

■ What does this look like?

Table 1: Effect of Conspiratorial Ideation and Institutional Trust on Predicted COVID-19 Mitigation Behaviors.

Conspiratorial Ideation	-0.518*** [0.062]	-0.465*** [0.086]				
Institutional Trust: State			0.182*** [0.047]	0.176*** (0.057)		
Institutional Trust: Federal					0.288*** [0.040]	0.301*** (0.051)
Female		0.196 [0.134]		0.088 (0.134)		0.134 (0.123)
Age		0.070 [0.088]		0.130 (0.092)		0.115 (0.084)
Income		-0.095 [0.084]		-0.069 (0.089)		-0.044 (0.082)
Education		-0.044 [0.110]		-0.037 (0.109)		-0.146 (0.101)
Ideology		-0.064 [0.042]		-0.155*** (0.049)		-0.058 (0.048)
Person of Color		-0.100 [0.182]		-0.091 (0.207)		0.015 (0.189)
COVID-19 Personal Experience		-0.055 [0.086]		-0.097 (0.097)		-0.045 (0.090)
Constant	1.345*** [0.155]	1.819*** [0.649]	-0.542*** [0.180]	0.395 (0.640)	-0.839*** [0.150]	-0.360 (0.605)
N	177	137	280	134	281	134
Adjusted R ²	0.250	0.211	0.060	0.101	0.184	0.242

Note. Values presented are linear regression estimates. Dependent variable is predicted COVID-19 behaviors - coded such that higher values indicate a higher likelihood of behaving in line with scientific recommendations for COVID-19 transmission mitigation. Gender and race are dummy variables coded such that 1 denotes female and person of color respectively. Standard errors in parentheses. Efron (1982) variant standard errors in brackets. *p<0.1; **p<0.05; ***p<0.01.

Linear Regression: Tables

Table 1: Effect of Conspiratorial Ideation and Institutional Trust on Predicted COVID-19 Mitigation Behaviors.

	Independent Variables			
Conspiratorial Ideation	-0.518*** [0.062]	-0.465*** [0.059]		
Institutional Trust: State		0.182*** [0.047]	0.176*** (0.057)	
Institutional Trust: Federal			0.288*** [0.040]	0.301*** (0.051)
Female	0.196 [0.134]	0.088 (0.134)		0.134 (0.123)
Age	0.070 [0.088]	0.130 (0.092)		0.115 (0.084)
Income	-0.095 [0.084]	-0.069 (0.089)		-0.044 (0.082)
Education	-0.044	-0.037		-0.146

Ideology		0.119 [0.118]		(0.109)	(0.101)	
Person of Color		-0.064 [0.041]		-0.155*** (0.049)	-0.058 (0.048)	
COVID-19 Personal Experience		-0.100 [0.182]		-0.091 (0.207)	0.015 (0.189)	
Constant	1.345*** [0.155]	1.419*** [0.649]	-0.542*** [0.180]	0.395 (0.640)	-0.839*** [0.150]	-0.360 (0.605)
N	177	177	280	134	281	134
Adjusted R ²	0.250	0.211	0.060	0.101	0.184	0.242

Note. Values presented are linear regression estimates. Dependent variable is predicted COVID-19 behaviors - coded such that higher values indicate a higher likelihood of behaving in line with scientific recommendations for COVID-19 transmission mitigation. Gender and race are dummy variables coded such that 1 denotes female and person of color respectively. Standard errors in parentheses. Efron (1982) variant standard errors in brackets. *p<0.1; **p<0.05; ***p<0.01.

Linear Regression: Multiple IVs

- We state the linear regression equation as:

$$y_i = \hat{\alpha} + \hat{\beta}x_i + \hat{u}_i$$

- It is important to note that this is functionally shorthand
- In the single IV case, x_i is a vector
- This equation works for multiple IVs as well

Linear Regression: Multiple IVs

- We state the linear regression equation as:

$$y_i = \hat{\alpha} + \hat{\beta}x_i + \hat{u}_i$$

- It is important to note that this is functionally shorthand
- In the single IV case, x_i is a vector
- This equation works for multiple IVs as well

$$y_i = \hat{\alpha} + \hat{\beta}_1x_{1i} + \hat{\beta}_2x_{2i} + \hat{\beta}_3x_{3i} + \cdots + \hat{u}_i$$

Linear Regression: Multiple IVs

y_i	$\hat{\beta}_1$	x_{1i}
1.2	1.2	1
2.4	1.2	2
3.6	1.2	3
4.8	1.2	4
6.0	1.2	5
7.2	1.2	6
8.4	1.2	7
9.6	1.2	8
10.8	1.2	9
12	1.2	10

Linear Regression: Multiple IVs

y_i	$\hat{\beta}_1$	x_{1i}	$\hat{\beta}_2$	x_{2i}
-0.8	1.2	1	-2	1
-1.6	1.2	2	-2	2
-2.4	1.2	3	-2	3
-3.2	1.2	4	-2	4
-4	1.2	5	-2	5
-4.8	1.2	6	-2	6
-5.6	1.2	7	-2	7
-6.4	1.2	8	-2	8
-7.2	1.2	9	-2	9
-8	1.2	10	-2	10

Linear Regression: Multiple IVs

y_i	$\hat{\beta}_1$	x_{1i}	$\hat{\beta}_2$	x_{2i}	$\hat{\beta}_3$	x_{3i}
-0.3	1.2	1	-2	1	0.5	1
-0.6	1.2	2	-2	2	0.5	2
-0.9	1.2	3	-2	3	0.5	3
-1.2	1.2	4	-2	4	0.5	4
-1.5	1.2	5	-2	5	0.5	5
-1.8	1.2	6	-2	6	0.5	6
-2.1	1.2	7	-2	7	0.5	7
-2.4	1.2	8	-2	8	0.5	8
-2.7	1.2	9	-2	9	0.5	9
-3	1.2	10	-2	10	0.5	10

Let's Try an Example Together

- Data from 2012 ANES
- Effect of SES on Party Identification
- Think about your data structure, and how this would apply as we go through this example

Linear Regression: Multiple IVs

- Multiple IVs complicate matters in two key ways
- First, the IVs may be correlated with each other violating one of the GM assumptions
- Second, the model is less capable of assigning the variance in outcomes due to one IV over another
- This issue increases exponentially, not linearly, with the addition of more and more Ivs

Linear Regression: Conclusion

- We know how to:
 - Calculate the linear best fit line (regression coefficient and constant)
 - Calculate uncertainty about the regression line
 - Calculate the coefficient of determination
 - **Interpret linear regression results**
- Remember: linear regression requires a continuous DV and a number of assumptions to function properly
- **With observational data, regression cannot make causal claims**

For Next Class

- Read the excerpt on iCollege for Thursday
- Complete Final Paper and submit by Thursday (7/28) by midnight